Overview of the TREC 2008 Enterprise Track

Krisztian Balog
ISLA, University of Amsterdam
k.balog@uva.nl

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

Paul Thomas
CSIRO, Australia
paul.thomas@csiro.au

Peter Bailey
Microsoft, USA
pbailey@microsoft.com

Nick Craswell
MSR Cambridge, UK
nickcr@microsoft.com

Arjen P. de Vries
CWI, The Netherlands
arjen@acm.org

1 Introduction

The goal of the enterprise track is to conduct experiments with enterprise data that reflect the experiences of users in real organizations. This year, we continued with the CERC collection introduced in TREC 2007 (Bailey et al., 2007). Topics were developed in conjunction with CSIRO Enquiries, who field email and telephone questions about CSIRO research from the public.

2 Collection

The CERC corpus (CSIRO Enterprise Research Collection, http://es.csiro.au/cerc/) represents the public-facing web of the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO). Here, we summarize the main characteristics of this corpus; a complete description of the collection is given by Bailey et al. (2007).

2.1 Data

The collection consists of all the *.csiro.au (public) websites as they appeared in March 2007. The resulting data set consists of 370 715 documents, with total size 4.2 gigabytes. The web crawler visited the outward-facing pages of CSIRO in a fashion similar to the crawl used in CSIRO's own search engine. In fact, the same crawler technology that CSIRO uses was used to gather the CSIRO documents (http://www.funnelback.com/). The corpus contains approximately 7.9 million hyperlinks, and 95% of pages have one or more outgoing links containing anchor text. One participant extracted email addresses of 3678 individuals, with 38% of documents containing at least one mailto field.

2.2 Users

When the CERC corpus was developed, a conscious decision was made to work with CSIRO employees to develop topics and make relevance judgments whenever possible. In 2007, this

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
NOV 2008

2. REPORT TYPE

3. DATES COVERED
00-00-2008 to 00-00-2008

4. TITLE AND SUBTITLE
Overview of the TREC 2008 Enterprise Track

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
National Institute of Standards & Technology (NIST),100 Bureau Drive,Gaithersburg,MD,20899-1070

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**12**	

role was filled by science communicators. Science communicators read and create the outward-facing web pages of CSIRO as part of their job to interact with industry, government agencies, professional groups, the media, and the public to promote the work of CSIRO.

This year, our users were staffers for CSIRO Enquiries. Enquiries staffers receive requests for information about CSIRO work, primarily via telephone and email. The “contact us” links on the bottom of most CERC pages lead to someone in Enquiries. Enquiries staffers need to search the CSIRO web to find the information needed to fulfill the request. Additionally, expert search could help them locate experienced CSIRO researchers to fill out gaps in what they find on the CSIRO web.

2.3 Tasks and Topics

The tasks this year are the same as in 2007: document search and expert search, although the goal of the user is somewhat different this year. An employee in CSIRO Enquiries is responding to an email request for information about something at CSIRO. To do this, they search the public-facing web for answers and resources. Additionally, they look for subject experts who can help them by providing in-depth information relating to the enquiry.

The topics have been extracted from a log of real email enquiries from January to March 2007, the same date range as the CERC crawl. They are not a random sample, but have been chosen to illustrate a range of requests. Each was answered with reference to at least one page on CSIRO’s public web site, so each has at least one relevant page in the corpus. There are a total of 77 topics, numbered CE-051 to CE-127.

Each topic has the original email (stripped of any identifying information, and of any greetings etc), and a short form which is a two- or three-word query created by track coordinator Paul Thomas but which Enquiries staff confirmed is very similar to one they’d issue to a search engine.

Here is an example topic:

```
<top>
<num>CE-053</num>
<query>selenium soil</query>
<narr>
Can you please provide a current e-mail address, or failing that can you
please put me in contact with the group responsible for the research into the
use of selenium as an additive to soils, to promote sheep
productivity/health. There were some trials conducted in WA and I am looking
for aditional information on these.
</narr>
</top>
```

The `query` field is the short form, as might be typed to a search engine; the `narr` field is the substantive part of the original email.

2.4 Assessments

This year, no CSIRO resources were available for making relevance judgments, so both document and expert search tasks were judged by participants.

Analysis of last years document judgment data indicated that participant judges needed additional resources if their judgments were to be comparable to those made by CSIRO “insiders” (Bailey et al., 2008). To that end, in addition to the topic text which includes the email sent to Enquiries, we provided judges with the final response sent by Enquiries, as well as a link to any CSIRO URL included in the response email. For expert search, the judges also received links to highly-relevant document search results for that same topic.

For the document search task, stratified sampling was used to select a subset of the pool to judge. The initial pool was the top 75 retrieved documents from two runs per group (selected according to priorities given by each group when the run was submitted). We then uniformly drew 100% of documents retrieved at ranks 1-3, 20% of documents ranked 4-25, and 10% of documents ranked 25-75. The principal measures for document search are mean inferred average precision ("infAP") and inferred NDCG ("infNDCG") (Yilmaz et al., 2008), which estimate AP and NDCG given the sample.

Topics were assigned to three different groups to study assessor effects. Participants judged the pools through the CSIRO assessment system (adapted from the assessment system used in the Million Query track).

The guidelines instructed the assessors to read the query and narrative, and optionally carry out a Web search to learn more about the subject. Relevance judgments were made on a three-point scale:

2: Highly likely to be a 'key page', containing an answer to the enquiry..
1: Possibly a 'key page'.
0: Not a 'key page', because, e.g., not relevant, off-topic, not an important page on the topic, on-topic but out-of-date, not the right kind of navigation point, or too informal or too narrow an audience.

For expert search, we drew a standard pool to a depth of 5 candidates from all submitted runs, along with the top 5 submitted supporting documents for each pooled candidate. The candidates and response email were compiled into an HTML file along with links to the supporting documents and highly relevant judged documents. Participants were asked to edit the file to indicate whether each candidate was or was not an expert. This simplified the process by not requiring an assessment platform (only some way to retrieve the linked documents), at the expense of some errors that may have crept in by editing the file by hand.

3 Document search task

For the document search tasks, participants were asked to return up to 1000 documents from the corpus in response to each topic. Each group was allowed to submit up to four runs. One run was required to be an automatic run using the `query` field.

Fourteen groups submitted a total of 56 document search runs. Of those, 39 were automatic runs using only the `query` field. 13 automatic runs used the `narr` field (email text) in addition to the `query`. There were four manual runs.

Figure 1 shows the range of infAP scores for all runs, ordered by mean infAP. The box for each run extends from the first to the third quartile of the infAP scores for each topic; the whiskers extend to include topics not more than 1.5 times the interquartile distance; outlier topics are shown as circles. The run names along the x-axis include "(n)" if the run used the `narr` field, and "(M)" indicates a manual run.

Table 1 shows the mean infAP and infNDCG scores for the top run from each group (by mean infAP). The top runs from each group were nearly always `query`-only automatic runs. Of the seven groups that submitted runs using the `narr` section in addition to the `query`, two groups — the University of Avingon and St. Petersburg State University (SPSU) — had `narr` runs perform better than their `query`-only runs. In all other cases, runs adding the `narr` field performed roughly the same, or otherwise much worse, than those using the `query` field only.

As a quick guide to papers by TREC participants appearing in the proceedings, we offer the following brief descriptions of each group's approaches.

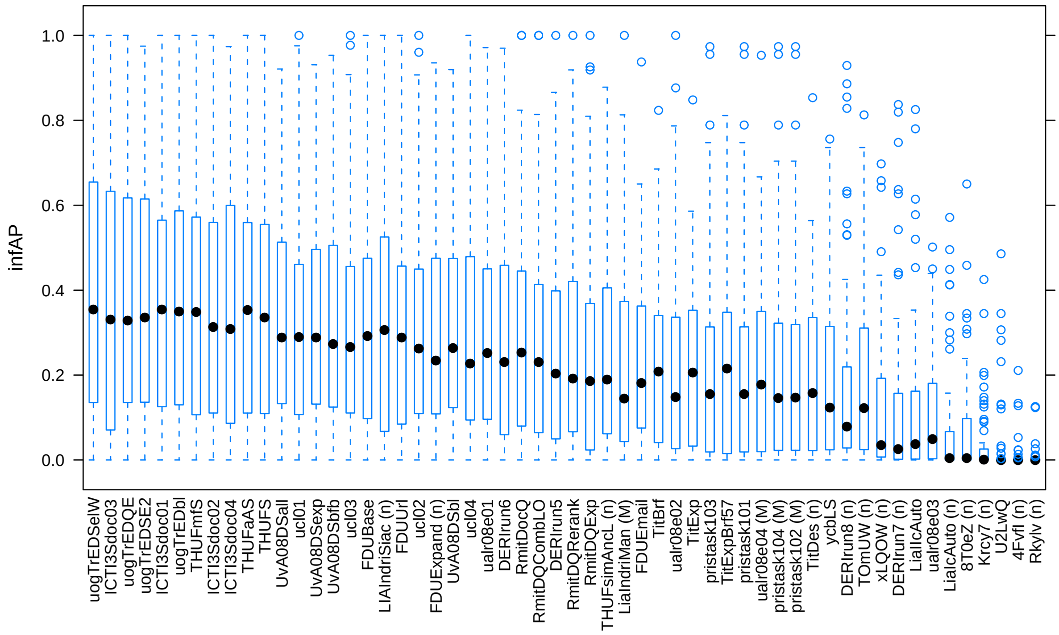

Figure 1: Box-and-whisker plots of infAP scores in the document search task, ordered by the run's mean infAP score across all topics. The box for each run extends from the first to the third quartile of the per-topic infAP scores; the whiskers extend to include topics not more than 1.5 times the interquartile distance; outliers are shown as circles.

Run	Group	Type	Fields	Mean infAP	infNDCG
uogTrEDSelW	UGlasgow	auto	q	0.3891	0.5660
ICTI3Sdoc03	CAS	auto	q	0.3760	0.5393
THUFmfS	Tsinghua	auto	q	0.3612	0.5578
UvA08DSall	UAmsterdam	auto	q	0.3306	0.4909
ucl01	UC-London	auto	q	0.3246	0.5175
FDUBase	Fudan	auto	q	0.3204	0.4985
LIAIndriSiac	UAvingon	auto	qn	0.3191	0.5078
ualr08e01	UArkansas	auto	q	0.3024	0.4838
DERIrun6	NUI-Galway	auto	q	0.3018	0.4791
RmitDocQ	RMIT	auto	q	0.2975	0.5045
TitBrf	Sebir	auto	q	0.2252	0.4035
pristask103	BUPT	auto	q	0.2216	0.4046
ycbLS	INRIA	auto	q	0.1879	0.3785
xLQOW	SPSU	auto	qn	0.1300	0.3057

Table 1: The top run from each group by mean infAP, showing the mean infAP and infNDCG scores for each.

UGlasgow looked at query expansion using external resources. The resources included blind feedback from web search engine results, and Wikipedia. (He et al., 2008)

CAS used BM25 and language models with blind feedback. (Shen et al., 2008)

Tsinghua investigated link analysis methods in the CERC collection, as well as selecting query-independent key pages based on outlinks and anchors. (Xue et al., 2008)

UAmsterdam developed a novel language model that mixes document models with expert profile models, as a collection enrichment technique. (Balog and de Rijke, 2008)

UC-London uses document search as one component of their expert search system. Their approach uses language models, and they investigate the use of anchor texts and in-degree counts. (Zhu, 2008)

(Fudan do not describe their document search approach in their paper.)

UAvingon tried a number of different approaches; their top run employs a passage retrieval method that comes from their work in QA.(SanJuan et al., 2008)

(UArkansas did not submit a final TREC proceedings paper as of this writing.)

NUI-Galway developed a term-weighting scheme based on BM25 that incorporates expert candidate profiles in determining the weights. (Cummins and O'Riordan, 2008)

RMIT investigated using out-degree of pages within the results list ("local outdegree") to rerank. (Wu et al., 2008)

Sebir investigated blind relevance feedback using Wikipedia as the expansion collection. (Peng and Mao, 2008)

(BUPT did not submit a final TREC proceedings paper as of this writing.)

INRIA investigated weighted PageRank variants, in particular first clustering the collection and differentially weighting links within and between clusters. (Nemirovsky and Avrachenkov, 2008)

SPSU looked at term and phrase weighting models based on entropy. (Nemirovsky and Dobrynin, 2008)

As stated above, each topic was assigned to three participant groups for relevance assessment. In the end, a total of 67 out of the full set of 77 topics were judged. 10 topics were judged by three groups, 33 topics by two groups, and 24 topics by only a single group. The first group assigned was labeled as the primary assessor, and the judgments of the primary assessor were used in the official results. Four topics (51, 74, 108, and 116) had no relevant documents judged by the primary assessor; these topics were not used in the official evaluation.

We also created two sets of relevance judgments using the other assessors. The first used the judgments of the second assessor, unless no such assessor existed, in which case the primary assessor's judgments were used. The second used the judgments of the third assessor where such existed (otherwise falling back to the second or primary assessor as available). We dropped the four topics where no relevant documents were judged by the primary assessor, as well as topic 63 which had no relevant documents judged by the secondary assessor. We computed mean infAP for all systems using these "secondary" and "tertiary" relevance judgments, and computed the Kendall's τ rank correlation between the order of systems by the official, secondary, and tertiary sets. The τ value was 0.92 between the official judgments and both the secondary and tertiary

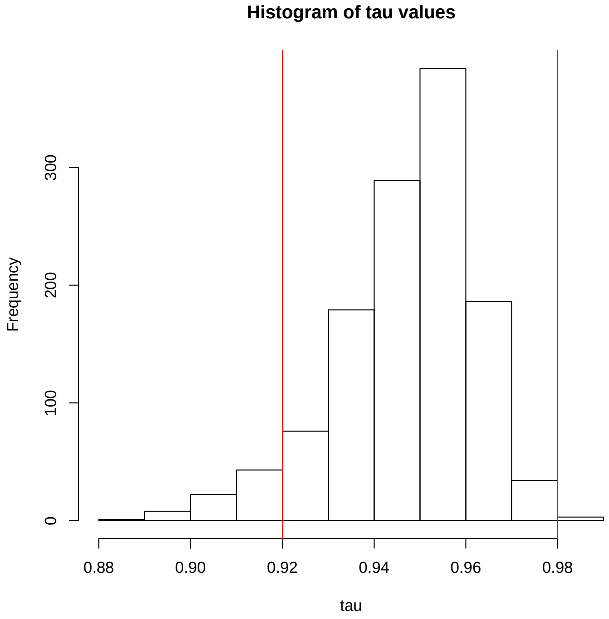

Figure 2: Distribution of τ values taken between all pairs of rankings in the randomly-selected-judge experiment. The vertical line at 0.92 is the τ between the primary judge and both the secondary and tertiary sets. The line at 0.98 is the τ between the secondary and tertiary sets.

(95% confidence interval 0.65–0.98 in both cases), and 0.98 between the secondary and tertiary rankings themselves (interval 0.75–0.999) (Kendall and Gibbons, 1990). We do not see any reason to believe there is a difference in correlations when judgement sets are changed.

Lastly, we also constructed fifty sets of relevance judgments choosing a judge (primary, secondary, or tertiary) at random for each topic, and compared the resulting rankings among each other. The lowest τ between two of these rankings was 0.89, and the highest was 0.99. The mean τ among all pairs of rankings was 0.95. Figure 2 shows the distribution of the τ values, along with the τs between the primary, secondary, and tertiary judgments for comparison. From this we conclude that although differences do exist among the relevance judgments, this does not have a large effect on the document search rankings.

4 Expert search task

For the expert search tasks, participants were asked to return email addresses of up to 100 candidate experts. Like in previous year, no canonical list of candidate experts was made available, email addresses were to be extracted from the data. Each group was allowed to submit up to four runs. Eleven groups submitted a total of 42 expert search runs. Of those, 32 were automatic runs using only the `query` field; 7 automatic runs used the `narr` field in addition to the `query`. Two groups submitted manual runs. Interestingly, the manual run `LiaIcExp08` by SanJuan et al. (2008) not only involved multiple iterations of manual query reformulations, but was created entirely manually by a paid search professional.

Table 2 shows the MAP and MRR scores for the top run from each group (by MAP). Below, we present a brief summary of participants' approaches.

Run	Group	Type	Fields	MAP	MRR
UvA08ESweb	UAmsterdam	auto	q	0.4490	0.8721
ICTI3Sexp01	CAS	auto	q	0.4214	0.7241
uogTrEXfeNPC	UGlasgow	auto	q	0.4126	0.7611
FDURoleRes	Fudan	auto	qn	0.4114	0.7516
THUPDDlchrS	Tsinghua	auto	q	0.3846	0.7419
WHU08NOPHR	Wuhan	auto	q	0.3826	0.6770
utqurl	UTwente	auto	q	0.3728	0.7647
UCLex04	UC-London	auto	q	0.3476	0.6759
DERIrun3	NUI-Galway	auto	q	0.2619	0.6212
LiaIcExp08	UAvingon	manual	qn	0.2513	0.8545
pristask204	BUPT	manual	qn	0.0977	0.2343

Table 2: The top run from each group by mean AP, showing the mean AP and mean RR scores for each. Reported results use the official qrels.

UAmsterdam used a combination of multiple approaches; a proximity-based version of their candidate model (Model 1B), the document-based model (Model 2), and a Web-based variation of Model 1B (to bring in external evidence). Additionally, they applied profile-based query expansion. (Balog and de Rijke, 2008)

CAS focused on identifying authoritative persons by constructing a recommendation network of persons, then applying the PageRank algorithm on this network. In addition, different weights were assigned to various types of person occurrences. (Shen et al., 2008)

UGlasgow applied a proximity-based variation of their Voting Model. They also investigated expanding candidate profiles with Web evidence. (He et al., 2008)

Fudan introduced two methods to judge whether a person is more likely to be an expert. One method is to determine the roles of a person by the context of pages; the other is to judge the authority of a person by exploiting the structure of specific document types. (Yao et al., 2008)

Tsinghua investigated the combination of profile-based and document-based methods. Link analysis and homepage detection were performed to identify high quality documents. They also experimented with automatic query type identification. (Xue et al., 2008)

Wuhan developed a model that considers the probability of query generation separately for different expert identifiers; the ambiguity of abbreviated person names was also addressed. Additionally, they adopted a method to detect phrases in the query. (Jiang et al., 2008)

UTwente combined the intranet-based ranking (produced using their infinite random walk based expert finding method) with various rankings obtained from the Web using search engine APIs. (Serdyukov et al., 2008)

UC-London uses a document-centric generative approach, and investigates the use of anchor texts and in-degree counts. Associations between candidates and query terms are captured using a combination of windows of different sizes. (Zhu, 2008)

NUI-Galway used genetic programming to find ranking functions, both for profiles-based and for document-based approaches. (Cummins and O'Riordan, 2008)

UAvingon carried out both automatic and manual search. The automatic method ranks summaries corresponding to email addresses using baseline Indri retrieval. The manual run employed multiple iterations of query refinement. (SanJuan et al., 2008)

(**BUPT** did not submit a final TREC proceedings paper as of this writing.)

			Ext.	Majority		Lenient		Unanimous		LiaIcExp08	
RunID	Type	Fields	res.	MAP	MRR	MAP	MRR	MAP	MRR	MAP	MRR
UvA08ESweb	auto	q	Y	**0.4490**	**0.8721**	**0.4199**	**0.8918**	**0.4921**	**0.7159**	0.2950	**0.4951**
UvA08EScomb	auto	q	N	0.4331	0.8547	0.4102	0.8855	0.4713	0.6966	0.2719	0.4510
ICTI3Sexp01	auto	q	N	0.4214	0.7241	0.3997	0.7462	0.4435	0.6169	0.2750	0.4393
ICTI3Sexp02	auto	q	N	0.4208	0.7275	0.4028	0.7549	0.4413	0.5956	0.2852	0.4574
ICTI3Sexp03	auto	q	N	0.4184	0.7243	0.4072	0.7612	0.4680	0.6302	0.2862	0.4388
uogTrEXfeNPC	auto	q	N	0.4126	0.7611	0.3991	0.7933	0.4372	0.6372	0.2710	0.4178
FDURoleRes	auto	qn	N	0.4114	0.7516	0.4005	0.8148	0.4430	0.6145	0.2930	0.4702
FDUExpRole	auto	qn	N	0.4112	0.7472	0.3989	0.8004	0.4404	0.6056	0.2838	0.4427
uogTrEXfePC	auto	q	N	0.3969	0.7259	0.3834	0.7671	0.4105	0.6013	0.2542	0.4027
UvA08ESm1b	auto	q	N	0.3935	0.8223	0.3696	0.8333	0.4411	0.6702	0.2556	0.4084
THUPDDlchrS	auto	q	N	0.3846	0.7419	0.3635	0.7806	0.4281	0.6025	**0.2987**	0.4923
WHU08NOPHR	auto	q	N	0.3826	0.6770	0.3924	0.7162	0.3740	0.5403	0.2513	0.3916
FDUExpRes	auto	qn	N	0.3815	0.6732	0.3848	0.7383	0.4086	0.5491	0.2468	0.3942
WHU08RFCAN	auto	q	N	0.3765	0.6884	0.3921	0.7605	0.3847	0.5568	0.2562	0.3769
uogTrEXmix	auto	q	Y	0.3749	0.7660	0.3536	0.8140	0.4167	0.6489	0.2467	0.4098
utqurl	auto	q	Y	0.3728	0.7647	0.3571	0.7816	0.4254	0.5965	0.2426	0.4216
FDUExpBase	auto	qn	N	0.3720	0.6430	0.3755	0.7070	0.4047	0.5570	0.2342	0.3746
utbase	auto	q	Y	0.3712	0.7399	0.3584	0.7689	0.4171	0.5796	0.2443	0.4160
utqtitle	auto	q	Y	0.3709	0.7541	0.3598	0.8012	0.4222	0.6059	0.2509	0.4399
THUPDDSlL	auto	qn	N	0.3707	0.7451	0.3488	0.7808	0.3881	0.5863	0.2986	0.4907
WHU08BASE	auto	q	N	0.3707	0.6563	0.3852	0.7157	0.4075	0.5738	0.2606	0.3843
utrecent	auto	q	Y	0.3701	0.7426	0.3543	0.7693	0.4145	0.5816	0.2546	0.4318
UvA08ESm2all	auto	q	N	0.3679	0.6831	0.3568	0.7482	0.3922	0.5442	0.2119	0.3265
THUPDDlcS	auto	q	N	0.3640	0.7176	0.3487	0.7461	0.3849	0.5713	0.2881	0.4585
WHU08CAN	auto	q	N	0.3609	0.6296	0.3753	0.7017	0.3539	0.5232	0.2163	0.3272
uogTrEXfeNP	auto	q	N	0.3535	0.7079	0.3463	0.7431	0.3554	0.5514	0.2255	0.3582
UCLex04	auto	q	N	0.3476	0.6759	0.3357	0.7117	0.3713	0.5289	0.2576	0.3927
THUPDDSwp	auto	q	N	0.3456	0.7485	0.3200	0.7691	0.3676	0.5664	0.2749	0.4592
UCLex03	auto	q	N	0.3433	0.6748	0.3328	0.7112	0.3751	0.5372	0.2586	0.3998
UCLex01	auto	q	N	0.3360	0.6789	0.3252	0.7152	0.3624	0.5318	0.2512	0.3796
UCLex02	auto	q	N	0.3346	0.6737	0.3261	0.7130	0.3550	0.5199	0.2540	0.3857
ICTI3Sexp04	auto	q	N	0.2860	0.6525	0.2693	0.7153	0.3218	0.5078	0.2519	0.4242
DERIrun3	auto	q	N	0.2619	0.6212	0.2621	0.6572	0.2670	0.4489	0.1797	0.2913
LiaIcExp08	manual	qn	Y	0.2513	0.8545	0.2163	0.8545	0.3513	0.6364	-	-
DERIrun2	auto	q	N	0.2164	0.6281	0.2032	0.6442	0.2425	0.4253	0.1602	0.2663
DERIrun1	auto	qn	N	0.1953	0.4706	0.1685	0.4923	0.1983	0.3031	0.1265	0.2086
LiaExp08	auto	q	N	0.1841	0.5502	0.1753	0.5801	0.1857	0.3666	0.1170	0.2278
DERIrun4	auto	qn	N	0.1758	0.4433	0.1705	0.4616	0.1932	0.3008	0.1060	0.1892
pristask204	manual	qn	N	0.0977	0.2343	0.1046	0.2709	0.1007	0.1624	0.0572	0.0820
pristask202	auto	q	N	0.0625	0.1332	0.0724	0.1915	0.0680	0.1125	0.0360	0.0634
pristask201	auto	q	N	0.0486	0.0999	0.0584	0.1621	0.0543	0.0787	0.0295	0.0515
pristask203	manual	qn	N	0.0476	0.1065	0.0578	0.1694	0.0480	0.0854	0.0277	0.0458

Table 3: All submitted runs, ordered by official MAP scores. MAP and MRR scores using different sets of qrels are also shown; highest scores for each are typeset in boldface.

Although the track did not solicit explicit baseline runs, we make some general observations of contrasting runs within several groups. Two groups (UAvignon and BUPT) submitted both automatic and manual runs; for both teams, their best performing submission was a manual run. Two groups (Tsinghua and NUI-Galway) had both `query`-only runs and runs using the `narr` field as well; the `query`-only runs performed better in both cases. Finally, two groups (UAmsterdam and UGlasgow) had submissions both with and without using external resources (Web search engine APIs). In one case (UAmsterdam) using external resources resulted in improvements, while in the other (UGlasgow) it did not.

As described in Section 2.4, relevance assessments were created by participants. Based on the judgments made, different sets of qrels could be created, depending on how agreement between assessors is handled. In addition, we also consider the manual run `LiaIcExp08` by SanJuan et al. (2008) as an alternative. Consequently, four different sets of relevance judgments were obtained; see Table 4.

Table 3 displays the MAP and MRR scores for all submitted runs, using the different sets of ground truth. Runs are ordered by their MAP scores according to the official set of qrels.

Qrels	Description	Avg. #experts per topic
Majority	A person is considered to be an expert if most assessors said so (tie votes taken as relevant). This was used as the official set of qrels.	10.4
Lenient	A person is considered to be an expert if at least one assessor said so.	12.6
Unanimous	A person is considered to be an expert if all assessors agreed.	4.8
LiaIcExp08	Judgments performed by an independent, external search professional (SanJuan et al., 2008).	2.4

Table 4: Alternative qrels sets for the expert finding task.

Qrels	Metric	Majority	Lenient	Unanimous	LiaIcExp08
Majority	MAP		0.8722	0.8420	0.5804
	MRR		0.9070	0.8072	0.6487
Lenient	MAP			0.7653	0.5560
	MRR			0.8257	0.6634
Unanimous	MAP				0.5926
	MRR				0.6243

Table 5: Kendall τ rank correlation.

To compare the rankings of systems using the different qrels sets we used Kendall's τ correlation. The systems defined by their runs, are ordered by some metric (MAP or MRR) for each qrels set, and the two rankings are compared. The run `LiaIcExp08` was ignored when using it as qrels. Table 5 reports the Kendall τ correlation given each qrels set against the other. We found strong correlation between the rankings of systems using the ground truths obtained from community judging (Majority, Lenient, and Unanimous). The `LiaIcExp08` qrels set showed moderate correlation against the others. One reason for that is that the number of experts identified for each topic is much lower than for the other qrels sets; in fact, according to the majority qrels, `LiaIcExp08` had the lowest recall of all runs. We also note that the professional's judgement is possibly more demanding than a participant's; and that the latter know how systems make ranking decisions and may themselves think similarly. We leave further examination and analysis to future work.

5 Summary

The fourth year of the enterprise track has featured the same tasks and collection as in the 2007 edition: document and expert search on the CERC corpus. Topics have been extracted from a log of real email enquiries. The only difference compared to the previous year is that both tasks were judged by participants. Although disagreements between assessors do exist, these do not have a large effect on the rankings of systems for either of the tasks.

Common themes for this year's document search task included query expansion using external sources (He et al., 2008; Peng and Mao, 2008), exploiting expertise profiles (Balog and de Rijke, 2008; Cummins and O'Riordan, 2008), and leveraging link-structure in the form of in-degree (Zhu, 2008), out-degree (Wu et al., 2008), or PageRank (Xue et al., 2008; Nemirovsky and Avrachenkov, 2008). The best performing document search run employed a query performance predictor mechanism to selectively apply collection enrichment (i.e., query expansion) based on Wikipedia on a per-query basis; retrieval was performed using the Divergence From Randomness framework (He et al., 2008).

As to expert search, methods and approaches employed this year included special treatment of different types of person occurrences (Shen et al., 2008; Yao et al., 2008; Jiang et al., 2008), link analysis (Xue et al., 2008; Zhu, 2008), proximity-based techniques (Balog and de Rijke, 2008; He et al., 2008; Zhu, 2008), the use of external evidence (Balog and de Rijke, 2008; He et al., 2008; Serdyukov et al., 2008), and the combination of candidate- and document-based methods (Balog and de Rijke, 2008; Xue et al., 2008). The best performing expert search run used a Language Modeling framework to combine three models: a proximity-based candidate model, a document-based model, and a Web-based variation of the candidate model (Balog and de Rijke, 2008).

	year			
Task	2005	2006	2007	2008
Expert search	9	23	15	11
E-mail known item search	18			
E-mail discussion search	14	10		
Document search			16	14

Table 6: Tasks and number of participating groups at the TREC Enterprise Track.

The Enterprise Track was introduced in 2005, and after four successful years, it came to an end in 2008. Since its introduction, the track, and especially the expert finding task, has generated a lot of interest within the research community, with rapid progress being made in terms of algorithms, modeling, and evaluation. Table 6 lists the tasks featured at the Enterprise track throughout the years. The Entity Search Track, implemented at TREC 2009 can be seen as a continuation of the expert search task, extending it along two dimensions: type (from people-only to multiple types of entities) and scale (from Intranet to Web).

References

P. Bailey, N. Craswell, I. Soboroff, and A.P. de Vries. The CSIRO enterprise search test collection. *SIGIR Forum*, 41(2):42–45, 2007.

P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A.P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter? In *Proceedings of the 31st Annual International ACM SIGIR Conference*, pages 667–674, Singapore, July 2008.

Krisztian Balog and Maarten de Rijke. Combining candidate and document models for expert search. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Ronan Cummins and Colm O'Riordan. DERI at TREC 2008 enterprise search track. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Ben He, Craig Macdonald, Iadh Ounis, Jie Peng, and Rodrygo L. T. Santos. University of Glasgow at TREC 2008: Experiments in blog, enterprise, and relevance feedback tracks with Terrier. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Jiepu Jiang, Wei Lu, and Haozhen Zhao. CSIR at TREC 2008 expert search task: Modeling expert evidence in expert search. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Maurice Kendall and Jean Dickinson Gibbons. *Rank correlation methods*. Oxford University Press, 5th edition, 1990.

Danil Nemirovsky and Konstantin Avrachenkov. Weighted PageRank: cluster-related weights. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Danil Nemirovsky and Vladimir Dobrynin. Word importance discrimination using context information. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Yefei Peng and Ming Mao. Blind relevance feedback with Wikipedia: Enterprise track. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Eric SanJuan, Nicolas Flavier, Fidelia Ibekwe-SanJuan, and Patrice Bellot. Universities of Avignon & Lyon III at TREC 2008: Enterprise track. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Pavel Serdyukov, Robin Aly, and Djoerd Hiemstra. University of Twente at the TREC 2008 enterprise track: Using the global web as an expertise evidence source. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Huawei Shen, Lei Wang, Wenjing Bi, Yue Liu, and Xueqi Cheng. Research on enterprise track of TREC 2008. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Mingfang Wu, Falk Scholer, and Steven Garcia. RMIT University at TREC 2008: Enterprise track. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Yufei Xue, Tong Zhu, Guichun Hua, Min Zhang, Yiqun Liu, and Shaoping Ma. THUIR at TREC2008: Enterprise track. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

Jing Yao, Jun Xu, and Junyu Niu. Using role determination and expert mining in the enterprise environment. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.

E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference*, pages 603–610, Singapore, July 2008.

Jianhan Zhu. The University College London at TREC 2008 enterprise track. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008.